Evaluation of ASR Systems using Information Retrieval Measures

Artem Peregooudou
IDIAP Research Institute, Martigny, Switzerland
Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland

ABSTRACT

Performance evaluation measures play an essential role in the design of automatic speech recognition systems as they are used to predict the performance of the system in a real application, to compare systems and to analyse the errors introduced during the recognition process. Commonly used word error rate (WER) has proven to be a good measure for most ‘old-school’ applications, such as dictation systems, where the whole content is of interest and every word has the same importance. However, with advances made in the recent past, more systems are currently emerging, aiming to resolve a new range of tasks. Multimodal systems, spoken document retrieval and call routing are examples of applications where the task involves categorisation and indexing of the audio content and where not all of the words have the same importance anymore. In this work we investigate on the evaluation of several ASR systems on the Ressource Management task using information retrieval measures.

DEFINITIONS / COMPARISON

Word Error Rate
both measures are based on an alignment between the reference transcription and the recogniser output

Information Retrieval Measures [1]

Indexed alignment:

- \( \text{WER} = \frac{S + D + I}{N} \)
  - \( N \): number of words in the reference transcription
  - \( S \): Substitution count
  - \( D \): Deletion count
  - \( I \): Insertion count

Well suited for an entire range of applications such as dictation systems.

Limitations:
- No upper bound
- Suffers from the error overestimation due to the dynamic programming alignment used for alignment [2].

References

EXPERIMENTS

Preliminary results obtained by evaluating 4 ASR systems on the DARPA Ressource Management task (continuous speech). 2 types of systems: HMM/GMM and TANDEM, based on 2 types of subword units: phonemes and graphemes. Currently, ongoing work on the evaluation of a call routing application.

IDF WEIGHTING

IDF (Inverse document frequency) is a measure showing how a word is discriminant with respect to a document set (in our case the documents are the different utterances).

For a word appearing in \( n \) documents out of a total of \( N \):

\[ \text{idf} = \log_2 \left( \frac{N}{n} \right) \]

<table>
<thead>
<tr>
<th>weighting scheme</th>
<th>all 1</th>
<th>all 1</th>
<th>all 1</th>
<th>all 1</th>
<th>all 1</th>
<th>all 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>IDF</td>
<td>IDF</td>
<td>IDF</td>
<td>IDF</td>
<td>IDF</td>
<td>IDF</td>
</tr>
<tr>
<td></td>
<td>GMMGraphemes</td>
<td>GMMPhonemes</td>
<td>XAVLGraphemes</td>
<td>TANDEMGMMPhonemes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>micro-IDF</td>
<td>0.2%</td>
<td>0.2%</td>
<td>94.0%</td>
<td>94.3%</td>
<td>94.1%</td>
<td>94.9%</td>
</tr>
<tr>
<td>macro-IDF</td>
<td>95.1%</td>
<td>92.6%</td>
<td>95.4%</td>
<td>94.3%</td>
<td>95.0%</td>
<td>95.1%</td>
</tr>
<tr>
<td>WBR (1-IDF)</td>
<td>95.1%</td>
<td>92.4%</td>
<td>95.6%</td>
<td>95.2%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

FUNCTION WORDS WEIGHTING

Function words obtained from the stop list used by the Idiap Text Retrieval system:
- 390 out of the 990 unique words of the vocabulary
- 4656 out of the 10288 words appearing in the utterances
- Function weights \( w_f \) varying from 0 to 1, non-function weights \( w_{\text{nf}} = 1 - w_f \)

F-measure against IDF for TANDEM Phoneme system

CONCLUSIONS / ONGOING & FUTURE WORK

- IR measures still suffers from the error underestimation due to the dynamic programming alignment. Looking for alternatives.
- Evaluation of call routing application.
- Evaluation of a multimodal ASR system

ACKNOWLEDGEMENTS

Big thanks to Matthew Magimai Doss (IDIAP) for his results on the RM task and his wise advices, to Jean-Yves Parfait (IDIAP) for his results on the call routing application, to John Dines (IDIAP) and Alessandro Vinciarelli (IDIAP) for the fruitful discussions we shared.